ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

-

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

UNIVERSITY OF ILLINOIS ENGINEERING EXPERIMENT STATION BULLETIN NO. 459

On the Solution of Ill-Conditioned, Simultaneous, Linear, Algebraic Equations by Machine Computation

by

B. T. Chao

H. L. Li

E. J. Scott

UNIVERSITY OF ILLINOIS BULLETIN

A REPORT OF AN INVESTIGATION

Conducted by

THE ENGINEERING EXPERIMENT STATION
UNIVERSITY OF ILLINOIS

Price: $1.00

Edited by R. Alan Kingery

UNIVERSITY OF ILLINOIS BULLETIN

Volume 58, Number 63; April, 1961. Published seven times each month by the University of Illinois. Entered as second-class matter December 11, 1912, at the post office at Urbana, Illinois, under the Act of August 24, 1912. Office of Publication, 49 Administration Building (West), Urbana, Ill.

On the Solution of Ill-Conditioned, Simultaneous, Linear, Algebraic Equations by Machine Computation

by

B. T. Chao

PROFESSOR OF MECHANICAL ENGINEERING

H. L. Li

RESEARCH ENGINEER, E. I. DU PONT DE NEMOURS AND COMPANY

E. J. Scott

ASSOCIATE PROFESSOR OF MATHEMATICS

ENGINEERING EXPERIMENT STATION BULLETIN NO. 459

ACKNOWLEDGMENTS

This work evolved from a project sponsored by the Office of Ordnance Research, United States Army, under contract No. DA-11-022-ORD-1980. A portion of the material appeared in Technical Report No. 7 submitted to that Office in June, 1959. The authors hereby express their appreciation to that office for the support of the program. Acknowledgment is made to Professor K. J. Trigger, Project Director, for his encouragement, and to Mrs. Donna Gargano for the typing of the manuscript.

2030—4-61—72741

UNIVERSITY
OF ILLINOIS
PRESS

CONTENTS

This page is intentionally blank.

I. INTRODUCTION

In the numerical analysis of many physical problems, oftentimes the formulation will lead to simultaneous, linear, algebraic equations. Field problems governed by Laplace, Poisson, and biharmonic equations are common examples. Numerical solutions of the diffusion equation using implicit representation offers another instance. In general, they find their application to the solution of various types of differential equations, both ordinary and partial. Due to the advent of modern high speed computers, seeking the solution of these equations becomes, in many cases, a daily routine even if the number of unknowns gets large.

The usual method for the solution of such linear equations is the Gauss elimination procedure. However, difficulties arise when the true solution is sensitive to round-off errors injected during computation. The latter may accumulate to such an extent that the numerical answers become completely worthless. Besides, the solution may *also* be sensitive to small errors in the coefficients and quantities at the right hand side. In either case, the determinant of the coefficient matrix is nearly singular. Such equations are described as ill-conditioned.

The occurrence of ill-conditioned equations in physical problems are by no means random. As pointed out by Turing,[1]* there is a large class of problems which naturally give rise to highly ill-conditioned systems. Head and Oulton[2] indicated their appearance in problems associated with aircraft design. In the evaluation of local temperatures at the sliding interface between a cutting tool and flowing metal chip, ill-conditioned systems were likewise encountered.[3] While the subject has received the attention of many people in recent years and some means of measuring "ill-conditionedness" have developed, one immediately discovers the lack of adequate information when attempting to solve such equations with high speed computers.

* Numbers in superscript refer to References Cited.

II. LITERATURE SURVEY

Attempts which have been made at solving ill-conditioned, simultaneous, linear, algebraic equations fall into two main categories. First, there is the indirect method, in which the unknowns are obtained by successive approximations. The second is the direct approach, in which the solution is obtained by a single application of the process.

An example of the indirect method is the widely publicized relaxation procedure of Southwell. Shaw,[4] Buckingham,[5] and Fox[6] gave detailed accounts of the procedure when applied to ill-conditioned systems. They all concluded that the convergence might become too slow to be of any practical value. In fact, to ensure convergence of the process, the coefficient matrix must be positive definite and symmetric.[7] While any matrix can be converted into a positive definite and symmetric form through multiplication by its transpose, the degree of ill-conditionedness is always aggravated as a consequence of such operation.[8] The relaxation procedure is also not suitable for machine computation.[9]

Another indirect method is the escalator process developed by Morris.[10] It consists of expressing the solution as a linear combination of column matrices or vectors which are orthogonal. A similar technique has been described by Fox, Huskey, and Wilkinson.[11] They are applicable only to symmetric matrices. The accuracy of Morris' method has been examined by Neville,[12] who criticized its lack of provision for estimating errors. Morris' procedure has also been criticized as not being well suited for automatic machine computation.[13]

Booth[13] proposed the method of steepest descent for solving ill-conditioned equations. The theory was based on positive definite and symmetric matrix form. Due to round-off errors, the actual path of solution may oscillate.

A unified convergence criterion valid both for iteration by total step (Jacobi) and iteration by single step (Gauss-Seidel) has been given by Geiringer.[14] Group iteration was also discussed. If the degree of ill-conditionedness were severe, the criterion would most probably not be met.

Direct methods of solution have also been examined. Buckingham[5] discussed the use of the Gauss elimination procedure and the difficulty of using it to solve ill-conditioned systems. Nielsen[15] favored Crout's modification of the Gauss elimination procedure. This modification will allow measurement of the degree of ill-conditionedness by the relative magnitude of the elements of the main diagonal as compared to the rest of the elements in the derived matrix. A re-arrangement of equations would be made if one of the diagonal elements is either very small or large. Bodewig[16] also discussed the possibility of suitably rearranging the rows and columns in order to reduce inherent errors in computation. Neither scheme is suitable for machine computation.

A method which combines iteration and elimination has also been suggested.[17] The approximate solution obtained from the computer is substituted into the original equations and the residuals calculated using double-precision computation. Corrections for the solutions are then calculated from these residuals. Forsythe[18] gives a survey of the many methods available for the solution of systems of linear, algebraic equations, and includes a rather extensive bibliography. This paper describes a new approach to the problem. It is condensed and modified from a technical report[19] to which two of the authors contributed. In many cases, improved results are obtained. Further work needs to be done to fully explore its merit.

III. A MEASURE OF ILL-CONDITIONEDNESS

In a two dimensional space, the ill-conditionedness of linear equations may be readily visualized geometrically. It is associated with the near "parallelism" of the straight lines represented by the equations. In an n-dimensional hyper-space, the linear equations may be interpreted as hyper-planes. The coefficients in the equations are proportional to the components of vectors normal to such planes. When the equations are ill-conditioned, two or more of these normals are nearly in the same direction.

Consider the set of simultaneous equations:

$$\sum_{j=1}^{n} a_{ij} x_j = f_i \quad (i = 1, 2, \ldots n) \qquad (3.1)$$

which in matrix form may be written:

$$AX = F \qquad (3.2)$$

where A is the coefficient matrix $[a_{ij}]$; X and F are column matrices. As pointed out by Booth,[13] the set (3.1) will be ill-conditioned if the absolute value of the determinant of the normalized coefficient matrix is very small as compared to unity. That is,

$$| A_N |_{abs} \ll 1 \qquad (3.3)$$

Other quantitative measures of ill-conditionedness have also been discussed in literature. Von Neumann and Goldstine[20] proposed the use of the so-called condition number, defined by $\frac{\lambda_{max}}{\lambda_{min}}$, where λ_{max} and λ_{min} are respectively the largest and smallest eigenvalues of the coefficient matrix. The larger this number, the higher the degree of ill-conditionedness is. Unfortunately, such a criterion is of little practical value, since the numerical determination of λ_{max} and λ_{min} may oftentimes be as long a process as the evaluation of the solutions of the original equations.[21]

Employing the concept of the condition number described above, Taussky[8] developed a useful theorem concerning ill-conditioned matrices. If A represents a real, non-singular matrix, and A' its transpose, then their product AA' is more "ill-conditioned" than A. This has been confirmed by Hartree.[21] It seems that this important result has not received the attention it deserves.

Turing[1] suggested the use of other condition numbers which also involve the multiplication of the coefficient matrix by its transpose. In view of Taussky's finding, such use has been discarded in favor of the simple criterion suggested by Booth (Eq. 3.3).

While the inequality (3.3) has been widely used to ascertain ill-conditionedness, it is by no means a sufficient condition under all circumstances. This may be demonstrated by the following examples.

Consider the set:

$$\left.\begin{aligned} 4.011\, x_1 + 4.012\, x_2 &= 0.001 \\ 4.012\, x_1 + 4.014\, x_2 &= 0.002 \end{aligned}\right\} \qquad \text{(a)}$$

which has the exact solution, $x_1 = -1$ and $x_2 = +1$. The solution as obtained from the ILLIAC,[1] using the method of elimination and with round-off at the 12th significant figure, has been found to be: $x_1 = -1.00000001819$ and $x_2 = +1.00000001819$. If errors of ± 3 to 10 hundredths of one percent are injected into the coefficients and quantities on the right hand side, the system will read:

$$\left.\begin{aligned} 4.012\, x_1 + 4.011\, x_2 &= 0.001001 \\ 4.013\, x_1 + 4.013\, x_2 &= 0.002002 \end{aligned}\right\} \qquad \text{(b)}$$

The corresponding ILLIAC solution is $x_1 = -1.0000022347$ and $x_2 = +1.0005011134$. These agree with the exact solution to a high degree of accuracy. On the other hand, when one computes $| A_N |$ of (a), it is found to be equal to 0.000705; very small indeed as compared to unity.

Let us now consider a different set which has the same coefficient matrix as that of (a). Thus,

$$\left.\begin{aligned} 4.011\, x_1 + 4.012\, x_2 &= 4010 \\ 4.012\, x_1 + 4.014\, x_2 &= 4010 \end{aligned}\right\} \qquad \text{(c)}$$

[1] The digital computer at the University of Illinois.

which has the exact solution, $x_1 = +2000$ and $x_2 = -1000$. The ILLIAC solution reads $x_1 = +2000.00003641$ and $x_2 = -1000.00003638$. Again, if errors of ± 3 to 10 hundredths of one percent are injected, we obtain

$$\left.\begin{aligned} 4.012\,x_1 + 4.011\,x_2 &= 4011 \\ 4.013\,x_1 + 4.013\,x_2 &= 4012 \end{aligned}\right\} \quad \text{(d)}$$

The solution as obtained from the ILLIAC now becomes: $x_1 = +999.5016083$ and $x_2 = +0.0002492015$. These in no way resemble the actual solution. Lastly, if errors are only introduced into the coefficients, i.e., if we seek the solution of the set:

$$\left.\begin{aligned} 4.012\,x_1 + 4.011\,x_2 &= 4010 \\ 4.013\,x_1 + 4.013\,x_2 &= 4010 \end{aligned}\right\} \quad \text{(e)}$$

we obtain: $x_1 = +1998.5048432$ and $x_2 = -999.2524135$. The foregoing examples clearly illustrate the inadequacy of describing the degree of ill-conditionedness by $|A_N|$ alone, without any reference to the quantities on the right hand side, namely, the f_i's.

As has been mentioned earlier, the ill-conditionedness of (a) is geometrically associated with the near parallelism of the lines represented by the equations. Changes in the quantities f_i's will execute a parallel shift of these lines. A question then naturally arises: Why is the ill-conditionedness of the set (a) not significantly influenced by the changes in f_i's as demonstrated, while the set (c) is? The reason becomes obvious if one observes that, in the latter case, the two f_i's are equal, and consequently, in the process of computation, one encounters the difficulty of taking small differences of two large and nearly identical numbers.

A mathematically rigorous analysis of ascertaining the ill-conditionedness of any set of linear equations needs separate research.

IV. A NEW ITERATIVE PROCEDURE

We shall now describe how to modify a given ill-conditioned matrix in order to make it a better conditioned one. Then we shall apply a new iterative procedure to the modified system to obtain a solution. It should be noted that the method does not require that the matrix be symmetric, positive definite, or both.

Consider the ill-conditioned normalized matrix of real numbers

$$A_N = [\bar{a}_{ij}] \tag{4.1}$$

Let us add to A_N the diagonal matrix

$$\Gamma = \begin{bmatrix} \gamma_1 & 0 & \cdots & 0 \\ 0 & \gamma_2 & \cdots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & \cdots & \gamma_n \end{bmatrix} \tag{4.2}$$

where the γ_i's are, for the moment, arbitrary real numbers. Let us now form the *modified matrix*[2]

$$A_N + \Gamma = \begin{bmatrix} \bar{a}_{11}+\gamma_1 & \bar{a}_{12} & \cdots & \bar{a}_{1n} \\ \bar{a}_{21} & \bar{a}_{22}+\gamma_2 & \cdots & \bar{a}_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \bar{a}_{n1} & \bar{a}_{n2} & \cdots & \bar{a}_{nn}+\gamma_n \end{bmatrix} \tag{4.3}$$

A measure of the improvement in the 'condition' of a system whose matrix of coefficients is given by (4.3) may be expressed by the ratio of the absolute values of the determinants of A_N and $A_N+\Gamma$, namely,

$$\beta = \frac{|A_N|_{abs}}{|A_N + \Gamma|_{abs}}$$

For, clearly, if $\beta \ll 1$, the improvement will be significant and the originally ill-conditioned matrix will become a well-conditioned one. β is called the *conditioning index.*

The problem now is one of selecting the γ_i's appropriately so that

$$|A_N + \Gamma|_{abs} > |A_N|_{abs} \tag{4.4}$$

Unfortunately, this problem is extremely involved and, instead, we solve a simpler one which in practice is oftentimes adequate. We take $\gamma_1 = \gamma_2 = \ldots = \gamma_n = \gamma$ and show that γ, under conditions stated below, can *always be chosen* so that (4.4) obtains.

For this case of equal γ_i's (4.3) becomes $A_N+\gamma I$, where I is the unit matrix, and, as is well known,

$$|A_N + \gamma I|_{abs} = |\gamma^n + p_1\gamma^{n-1} + p_2\gamma^{n-2} + \ldots + p_{n-1}\gamma + p_n| \equiv |P(\gamma)|, \tag{4.5}$$

wherein the p_i's are constants depending only on the elements $\bar{a}_{ij}$ and the vertical bars on the right side of the equal sign are now absolute value signs. Clearly,

$$|P(0)| \equiv |p_n| = |A_N|_{abs} \neq 0$$

since we are considering non-singular matrices. Therefore, if $p_{n-1} \neq 0$, there will be no relative maximum or minimum of $P(\gamma)$ at $\gamma = 0$. Hence, in view of the continuity of $P(\gamma)$, there exists a $\sigma > 0$, such that

$$|P(\gamma)|_{\gamma=\sigma} > |A_N|_{abs}, \tag{4.6}$$

or

$$|P(\gamma)|_{\gamma=-\sigma} > |A_N|_{abs} \tag{4.7}$$

In practice, one computes the determinants $|A_N+\gamma I|_{abs}$ for $\gamma = \pm\sigma$ and selects the larger of the two quantities.

If $p_{n-1}=0$, which corresponds to a rare matrix form, $P(\gamma)$ may have a relative maximum or minimum, or an inflection point at $\gamma = 0$. If a *maximum* occurs, it is not possible to improve the matrix by the present method.

Let us suppose that the ill-conditioned system of equations

$$\sum_{j=1}^{n} a_{ij} x_j = f_i \quad (i = 1, 2, \ldots, n) \tag{4.8}$$

is modified according to the present method. Then the improved system will read:

[2] In a private communication, Dr. S. D. Conte of Space Technology Laboratories, California, recently informed the authors that the proposed method was similar to the one used by Riley.[22] However, Riley considered only the positive definite and symmetric matrix.

$$\begin{aligned}
&(a_{11}+\gamma_1)\,x_1+a_{12}\,x_2+\ldots+a_{1n}\,x_n=f_1+\gamma_1\,x_1\\
&a_{21}\,x_1+(a_{22}+\gamma_2)\,x_2+\ldots+a_{2n}\,x_n=f_2+\gamma_2\,x_2\\
&\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\\
&a_{n1}\,x_1+a_{n2}\,x_2+\ldots+(a_{nn}+\gamma_n)\,x_n=f_n+\gamma_n\,x_n
\end{aligned}\tag{4.9}$$

Note: In what follows the restriction of equal γ_i's will not be imposed because the analysis is perfectly general and does not require it.

Mathematically speaking, equations (4.8) and (4.9) are equivalent. However, their behavior with respect to arithmetical operations in machine computation may be entirely different. This arises from the fact that the coefficient matrix in (4.9) is better conditioned.

To solve the system (4.9) we shall adopt the following iterative procedure. In the first step we delete the terms $\gamma_i x_i (i=1, 2, \ldots, n)$ on the right side of equations (4.9) and solve the resulting system on the ILLIAC. The solution so obtained we designate by

$$\Xi^{(1)} = \begin{bmatrix} \xi_1^{(1)} \\ \xi_2^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ \xi_n^{(1)} \end{bmatrix} \tag{4.10}$$

If the exact solution of (4.8) is X, the difference $X-\Xi^{(1)}$ will then be the error in the first iterative solution and, accordingly, we write

$$E^{(1)} = X - \Xi^{(1)} = \begin{bmatrix} e_1^{(1)} \\ e_2^{(1)} \\ \cdot \\ \cdot \\ \cdot \\ e_n^{(1)} \end{bmatrix} \tag{4.11}$$

By comparing the original set of equations (4.8) with the set satisfied by $\Xi^{(1)}$, we obtain

$$\begin{aligned}
&a_{11}\,e_1^{(1)}+a_{12}\,e_2^{(1)}+\ldots+a_{1n}\,e_n^{(1)}=\gamma_1\,\xi_1^{(1)}\\
&a_{21}\,e_1^{(1)}+a_{22}\,e_2^{(1)}+\ldots+a_{2n}\,e_n^{(1)}=\gamma_2\,\xi_2^{(1)}\\
&\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\quad\cdot\\
&a_{n1}\,e_1^{(1)}+a_{n2}\,e_2^{(1)}+\ldots+a_{nn}\,e_n^{(1)}=\gamma_n\,\xi_n^{(1)}
\end{aligned}\tag{4.12}$$

or in matrix notation

$$AE^{(1)} = \Gamma\Xi^{(1)} \tag{4.13}$$

Geometrically, we may interpret this as a translation of coordinates from X to $E^{(1)}$. We observe that the coefficient matrix of (4.12) is the same as that of (4.8), and hence it cannot be solved directly on the computer. To solve (4.13) we simply proceed as before, and write

$$(A+\Gamma)\,\Xi^{(2)} = \Gamma\Xi^{(1)} \tag{4.14}$$

where again the term on the right side, namely, $\Gamma\Xi^{(2)}$ has been deleted. The system (4.14) is again solved by the ILLIAC. Let us denote this solution by

$$\Xi^{(2)} = \begin{bmatrix} \xi_1^{(2)} \\ \xi_2^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ \xi_n^{(2)} \end{bmatrix} \tag{4.15}$$

The difference $E^{(1)}-\Xi^{(2)}$ is the error incurred in calculating $\Xi^{(2)}$ and we designate it by $E^{(2)}$. Therefore, $E^{(2)}=E^{(1)}-\Xi^{(2)}=X-\Xi^{(1)}-\Xi^{(2)}$. If the foregoing procedure is repeated m times, we obtain

$$E^{(m)} = E^{(m-1)} - \Xi^{(m)} = X - \sum_{j=1}^{m} \Xi^{(j)}$$

or

$$X = \sum_{j=1}^{m} \Xi^{(j)} + E^{(m)} \tag{4.16}$$

If the method converges, then the solution of the original system is

$$X = \sum_{j=1}^{\infty} \Xi^{(j)} \tag{4.17}$$

The condition under which convergence takes place and the expression for the error term after m iterations will be given later.

At this point we list the formulae for the determination of $\Xi^{(j)}$'s, namely,

$$(A+\Gamma)\,\Xi^{(m)} = \begin{cases} F & \text{when } m=1, \\ \Gamma\Xi^{(m-1)} & \text{when } m>1. \end{cases} \tag{4.18}$$

An important feature of this method is that the coefficient matrix for the determination of successive iterations remains the same. This fact makes it possible for the complete iterative computation to be readily programmed on the ILLIAC.

As has been pointed out previously, in the present method of calculation there is a displacement in the origin of the coordinate axes for the unknowns in the equations after every cycle of iteration. Errors will thus likely be accumulated

in the sum $\sum_{j=1}^{r} \Xi^{(j)}$ as the process continues. This source of error may be conveniently removed by starting anew with (4.9) after the r-th cycle, but this time with the terms $\gamma_i x_i$ $(i=1, 2, \ldots, n)$ at the right hand side of the equations not set equal to zero, but replaced by the terms appearing in the column matrix $\Gamma \sum_{j=1}^{r} \Xi^{(j)}$. The iteration proceeds in the usual manner and this artifice may be repeated if necessary.

A. THE GEOMETRY OF THE NEW ITERATIVE PROCESS

Before discussing the convergence criterion, it is interesting and instructive to describe geometrically the nature of the convergence of the new procedure. For ease of illustration we shall consider a system of two equations in two unknowns. Suppose, therefore, that the given system is

$$\left.\begin{aligned} a_{11}\, x_1 + a_{12}\, x_2 &= f_1 \\ a_{21}\, x_1 + a_{22}\, x_2 &= f_2 \end{aligned}\right\} \qquad \begin{aligned} &\text{(i)} \\ &\text{(ii)} \end{aligned}$$

Modifying it for the first iteration, we have

$$\begin{aligned} (a_{11} + \gamma_1)\, \xi_1^{(1)} + a_{12}\, \xi_2^{(1)} &= f_1, && \text{(i)}^{(1)} \\ a_{21}\, \xi_1^{(1)} + (a_{22} + \gamma_2)\, \xi_2^{(1)} &= f_2, && \text{(ii)}^{(1)} \end{aligned}$$

in which γ_1 and γ_2 are of the same sign as that of a_{11} and a_{22}. In Fig. 1, the original equations are

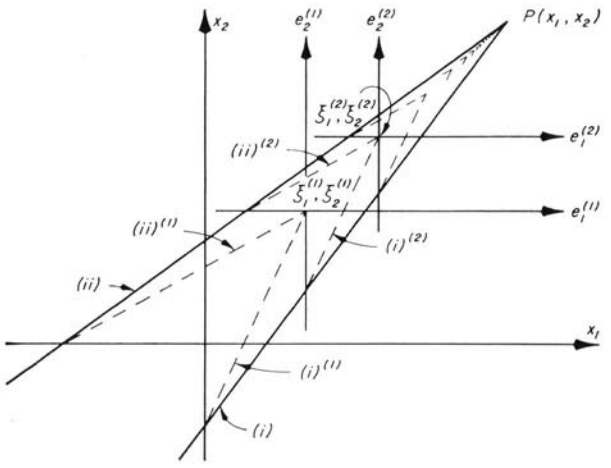

Fig. 1

represented by lines (i) and (ii), while the modified set is represented by lines $(i)^{(1)}$ and $(ii)^{(1)}$. The latter intersect at a larger angle and hence they are better conditioned. Their point of intersection $(\xi_1^{(1)}, \xi_2^{(1)})$ represents the result of the first iteration as well as the origin of the new coordinate system $e_1^{(1)}$ and $e_2^{(1)}$. The original equations, when referred to the new coordinates, are:

$$\begin{aligned} a_{11}\, e_1^{(1)} + a_{12}\, e_2^{(1)} &= \gamma_1\, \xi_1^{(1)}, \\ a_{21}\, e_1^{(1)} + a_{22}\, e_2^{(1)} &= \gamma_2\, \xi_2^{(1)}. \end{aligned}$$

For the second iteration, the modified equations become:

$$\begin{aligned} (a_{11} + \gamma_1)\, \xi_1^{(2)} + a_{12}\, \xi_2^{(2)} &= \gamma_1\, \xi_1^{(1)} && \text{(i)}^{(2)} \\ a_{21}\, \xi_1^{(2)} + (a_{22} + \gamma_2)\, \xi_2^{(2)} &= \gamma_2\, \xi_2^{(1)}, && \text{(ii)}^{(2)} \end{aligned}$$

which are represented by lines $(i)^{(2)}$ and $(ii)^{(2)}$. Their intersection is the point $(\xi_1^{(2)}, \xi_2^{(2)})$. Provided the method yields a convergent sequence, then, by repeating the foregoing procedure a sufficient number of times, the successive intersections tend to the required solution (x_1, x_2).

B. THE CONVERGENCE CRITERION AND ERROR ESTIMATION

On the basis of equations (4.18), we have

$$\begin{aligned} &\Xi^{(1)} = (A + \Gamma)^{-1} F \equiv C = [c_{ij}] \\ &\text{and} \\ &\Xi^{(m)} = (A + \Gamma)^{-1}\, \Gamma \Xi^{(m-1)}, \quad m > 1 \end{aligned} \qquad \textbf{(4.19)}$$

Let $(A+\Gamma)^{-1}\Gamma \equiv B = [b_{ij}]$. Then for $m>1$, we have

$$\begin{aligned} \Xi^{(m)} &= B\Xi^{(m-1)} \\ \Xi^{(m-1)} &= B\Xi^{(m-2)} \\ &\cdots \\ \Xi^{(3)} &= B\Xi^{(2)} \\ \Xi^{(2)} &= B\Xi^{(1)} = BC \end{aligned}$$

Therefore,

$$\Xi^{(m)} = B^{m-1}C, \quad m \geqslant 1$$

where $B^0 = I$, and

$$\sum_{j=1}^{m} \Xi^{(j)} = (I + B + B^2 + \ldots + B^{m-1})\, C.$$

Let us now assume that $\sum_{j=0}^{\infty} B^j$ converges. Then, in view of the fact that $X = C + BX$, we have

$$X = (I - B)^{-1} C = \sum_{j=0}^{\infty} B^j C,$$

and therefore

$$X - \sum_{j=1}^{m} \Xi^{(j)} = \left(\sum_{j=0}^{\infty} B^j - \sum_{j=o}^{m-1} B^j \right) C$$

$$= \sum_{j=m}^{\infty} B^j C = (I - B)^{-1} B^m C$$

$$= B^m (I - B)^{-1} C \equiv E^{(m)}. \quad \textbf{(4.20)}$$

Now, $\lim_{m \to \infty} B^m = 0$ because $\sum_{j=0}^{\infty} B^j < \infty$.
Therefore, taking the limit of both sides of (4.20) as $m \to \infty$, we see that $\lim_{m \to \infty} E^{(m)} = 0$
and

$$X = \sum_{j=1}^{\infty} \Xi^{(j)}$$

Consequently, we have the following **Theorem. If $\sum_{j=o}^{\infty} B^j$ converges, then the series $\sum_{j=1}^{\infty} \Xi^{(j)}$ converges to the solution X. Furthermore, the error $E^{(m)}$ committed after m iterations is given by either $(I-B)^{-1} B^m C$ or $B^m (I-B)^{-1} C$.**

In summary, it should be emphasized that the method here discussed requires the satisfaction of two conditions, namely, (i) the improvement of the condition of the matrix and (ii) the convergence of $\sum_{j=o}^{\infty} B^j$. The selection of Γ (or even γI) which will ensure the convergence of $\sum_{j=o}^{\infty} B^j$ is a difficult problem. Even after this is done the computation of $E^{(m)}$ can be formidable. Consequently, we propose the following simple alternative; and we seek an expression for an upper bound of error instead of calculating the error itself. In practice, this is what one usually needs.

Referring to (4.19), let us denote the inverse of the matrix $(A+\Gamma)$ by D. Thus,

$$(A + \Gamma)^{-1} \equiv D = [d_{ij}]. \quad \textbf{(4.21)}$$

Then the r-th row of $\Xi^{(m)}$ $(m>1)$ will be given by

$$\xi_r^{(m)} = d_{r1} \gamma_1 \xi_1^{(m-1)} + d_{r2} \gamma_2 \xi_2^{(m-1)} + \ldots + d_{rn} \gamma_n \xi_n^{(m-1)},$$
$$r = 1, 2, 3, \ldots, n,$$

for which the following inequality holds,

$$\begin{aligned} \left| \xi_r^{(m)} \right| &\leqslant \left| d_{r1} \gamma_1 \xi_1^{(m-1)} \right| + \left| d_{r2} \gamma_2 \xi_2^{(m-1)} \right| + \ldots + \left| d_{rn} \gamma_n \xi_n^{(m-1)} \right| \\ &\leqslant \left| \gamma_M \right| \left\{ \left| d_{r1} \right| + \left| d_{r2} \right| + \ldots + \left| d_{rn} \right| \right\} \left| \xi_M^{(m-1)} \right|, \end{aligned} \quad \textbf{(4.22)}$$

where $|\gamma_M|$ and $|\xi_M^{(m-1)}|$ are the maximum values of the $|\gamma_i|$'s and the $|\xi_i^{(m-1)}|$'s $(i=1, 2, 3, \ldots, n)$, respectively. Since (4.22) holds for $r=1, 2, 3, \ldots, n$, it will hold if the left hand side is replaced by

$$\left| \xi_M^{(m)} \right| = \max \left\{ \left| \xi_1^{(m)} \right|, \left| \xi_2^{(m)} \right|, \ldots, \left| \xi_n^{(m)} \right| \right\}$$

Therefore, we may write

$$\left| \xi_M^{(m)} \right| \leqslant \left| \gamma_M \right| \left\{ \sum_{i=1}^{n} \left| d_{ri} \right| \right\} \left| \xi_M^{(m-1)} \right|,$$
$$r = 1, 2, 3, \ldots, n. \quad \textbf{(4.23)}$$

Let $d_M = \max \left\{ \sum_{i=1}^{n} \left| d_{1i} \right|, \sum_{i=1}^{n} \left| d_{2i} \right|, \ldots, \sum_{i=1}^{n} \left| d_{ni} \right| \right\}$,

and let us require that the product

$$\left| \gamma_M \right| d_M = K < 1, \quad \textbf{(4.24)}$$

or, in the case of equal γ_i's,

$$\left| \gamma \right| d_M = K < 1. \quad \textbf{(4.25)}$$

K will be referred to as the *convergence constant.*

These conditions *can always be satisfied* by choosing γ_M or γ sufficiently small. To be sure, small values of γ may result in only a slight improvement of the matrix condition. Consequently, a compromise has to be made which depends on the degree of ill-conditionedness of the original set of equations, the round-off errors injected during the computation, cost of machine time, and the accuracy required in the final result.

Returning to (4.22) and (4.24), we have for $m>1$,

$$\left| \frac{\xi_M^{(m)}}{\xi_M^{(m-1)}} \right| \leqslant K < 1, \quad \textbf{(4.26)}$$

where it is assumed that $\xi_M^{(m-1)} \neq 0$. Therefore, by a well-known theorem, the series $\sum_{m=1}^{\infty} \xi_M^{(m)}$ is absolutely convergent.

From the foregoing discussion it is seen that, in some instances, it is possible to relax the restriction imposed by (4.24) or (4.25), i.e., to allow $|\gamma| d_M$ to go slightly beyond unity, yet (4.26) is still satisfied. Example 2 in the following section is selected to demonstrate this point.

Consider now the expression for each individual component of the matrix X, namely,

$$x_r = \sum_{j=1}^{\infty} \xi_r^{(j)} = \sum_{j=1}^{m} \xi_r^{(j)} + \sum_{j=m+1}^{\infty} \xi_r^{(j)}, \; r = 1, 2, 3, \ldots, n.$$

The error committed after m iterations is given by

$$e_r^{(m)} \equiv x_r - \sum_{j=1}^{m} \xi_r^{(j)} = \sum_{j=m+1}^{\infty} \xi_r^{(j)}, \quad r = 1, 2, 3, \ldots, n,$$

and

$$\left| e_r^{(m)} \right| < \sum_{j=m+1}^{\infty} \left| \xi_r^{(j)} \right|, \quad r = 1, 2, 3, \ldots, n.$$

However, in view of the fact that $|\xi_r^{(j)}| \leqslant |\xi_M^{(j)}|$ and that (4.26) obtains, we have

$$\begin{aligned} \left| e_r^{(m)} \right| &< K \left| \xi_M^{(m)} \right| (1 + K + K^2 + \ldots) \\ &= \frac{K}{1 - K} \left| \xi_M^{(m)} \right|, \end{aligned} \tag{4.27}$$

$r = 1$ or 2 or 3 . . . or n. Thus, the absolute values of the errors committed after m iterations are bounded by $|\xi_M^{(m)}| \cdot K/(1-K)$.

In conclusion, the authors wish to emphasize the fact that the procedure proposed in the paper does not exclude the use of any refined programs of computer calculation, such as the double precision routine, when the modified set (4.18) is solved. On the other hand, such refined computation procedure is *not* a substitute for the present method, which hinges on an appropriate modification of the coefficient matrix and iteration with progressive displacements of the vector Ξ. Rather it supplements the many known methods of solving linear systems when the latter are ill-conditioned.

V. NUMERICAL EXAMPLES

We shall now demonstrate the usefulness of the proposed method by two numerical examples. In each case, improvement on the solution was clearly indicated as compared with the well-known "method of elimination."

Example 1:

Consider the set:

$$\left.\begin{aligned} 4.011\,x_1 + 4.012\,x_2 &= 1.000 \\ 4.012\,x_1 + 4.014\,x_2 &= 2.100 \end{aligned}\right\} \qquad (5.1)$$

whose solution accurate to 10 significant figures is $x_1 = 1098.664654$ and $x_2 = +1098.640407$.

In order to examine the influence of round-off errors on the solution we arbitrarily decree that all calculations be carried out to 4 significant figures. Using the method of elimination, one obtains:

$$x_1 = -549.8$$
$$x_2 = +550.0$$

These represent errors of approximately 50%. Next, we solve (5.1) by the proposed iterative method. For $m = 1$, the modified equations become:

$$\left.\begin{aligned} (4.011 + \gamma)\,\xi_1^{(1)} + 4.012\,\xi_2^{(1)} &= 1.000 \\ 4.012\,\xi_1^{(1)} + (4.014 + \gamma)\,\xi_2^{(1)} &= 2.100 \end{aligned}\right\} \qquad (5.2a)$$

If γ is selected to be 0.002, one finds $\beta = 0.044$, which is much less than unity. The elements in the inverse matrix D as defined in (4.21) are:

$$\begin{bmatrix} 200.2 & -200.0 \\ -200.0 & -200.1 \end{bmatrix}$$

from which one computes the convergence constant:

$K = \gamma d_m = 0.002\{|200.2| + |-200.1|\} = 0.8006$, which is also less than unity. This ensures the convergence of the method. With $\gamma = 0.002$, and using also the method of elimination, (5.2a) has the solution:

$$\xi_1^{(1)} = -220.0$$

and

$$\xi_2^{(1)} = +220.0$$

For $m = 2$, the modified equations are:

$$\left.\begin{aligned} 4.013\,\xi_1^{(2)} + 4.012\,\xi_2^{(2)} &= \gamma\xi_1^{(1)} = -0.4400 \\ 4.012\,\xi_1^{(2)} + 4.016\,\xi_2^{(2)} &= \gamma\xi_2^{(1)} = +0.4400 \end{aligned}\right\} \qquad (5.2b)$$

whose solution is:

$$\xi_1^{(2)} = -175.9$$
$$\xi_2^{(2)} = +175.9$$

The process has been repeated and the results obtained from the first fifteen cycles of computation are tabulated below:

m	$\xi_1^{(m)}$	$\xi_2^{(m)}$
1	−220.0	220.0
2	−175.9	175.9
3	−140.7	140.7
4	−112.5	112.5
5	−90.00	90.00
6	−71.98	71.98
7	−57.60	57.60
8	−46.08	46.08
9	−36.86	36.86
10	−29.48	29.48
11	−23.58	23.58
12	−18.86	18.86
13	−15.09	15.09
14	−12.07	12.07
15	−9.656	9.656

Thus,

$$x_1 = \sum_{j=1}^{15} \xi_1^{(j)} + e_1^{(15)} = -1061 + e_1^{(15)}$$

and

$$x_2 = \sum_{j=1}^{15} \xi_2^{(j)} + e_2^{(15)} = +1061 + e_2^{(15)}$$

Using (4.27) one finds $|e_1^{(15)}|$ or $|e_2^{(15)}| < \frac{0.8006}{1-0.8006} \times |9.656| = 38.77$. This compares with actual errors of $|-1098.664654 + 1061| \simeq 37.66$ and $|1098.640407 - 1061| \simeq 37.64$ respectively.

Example 2:

Consider the severely ill-conditioned set:

$$\left.\begin{aligned} -3.0000000000\,x_1 + 2.9999999999\,x_2 \\ + 2.9999999999\,x_3 = 8.9999999998 \\ -2.9999999999\,x_1 + 3.0000000000\,x_2 \\ + 2.9999999999\,x_3 = 8.9999999998 \\ -2.9999999999\,x_1 + 2.9999999999\,x_2 \\ + 3.0000000000\,x_3 = 8.9999999998 \end{aligned}\right\} \quad (5.3)$$

which has the exact solution $x_1 = -1$, $x_2 = 1$ and $x_3 = 1$ as can be verified by direct substitution. The ILLIAC solution with round-off error at the 12th decimal place has been found to be:

$$\begin{aligned} x_1 &= -0.2857142857 \\ x_2 &= 1.2857142856 \\ x_3 &= 1.4285714287 \end{aligned}$$

In this case, if the convergence criterion (4.25) is to be satisfied, one has to select an extremely small value of γ. This would result in an insufficient improvement in the matrix condition. As pointed out earlier in the text, one may attempt to relax the restriction on γ and allow $|\gamma| d_M$ to assume a value slightly higher than unity. Let us thus try: $\gamma_1 = -0.1$, $\gamma_2 = \gamma_3 = 0.1$. The corresponding modified set is,

$$\left.\begin{aligned} -3.1000000000\,\xi_1^{(1)} + 2.9999999999\,\xi_2^{(1)} \\ + 2.9999999999\,\xi_3^{(1)} = 8.9999999998 \\ -2.9999999999\,\xi_1^{(1)} + 3.1000000000\,\xi_2^{(1)} \\ + 2.9999999999\,\xi_3^{(1)} = 8.9999999998 \\ -2.9999999999\,\xi_1^{(1)} + 2.9999999999\,\xi_2^{(1)} \\ + 3.1000000000\,\xi_3^{(1)} = 8.9999999998 \end{aligned}\right\} \quad (5.4)$$

When the elements in the inverse matrix D are evaluated and d_M computed therefrom, one finds that $|\gamma| d_M = 1.33$.

The numerical value of the conditioning index β can not be found since, due to the extreme ill-conditionedness of the original set, it is not possible to evaluate $|A_N|$ by the computer. This is immaterial, however, because it is obvious that $\beta \ll 1$.

At this stage of calculation, there is no assurance that the method will converge. We simply proceed with the iteration in the usual way and obtain the results of the first six cycles as follows:

m	$\xi_1^{(m)}$	$\xi_2^{(m)}$	$\xi_3^{(m)}$
1	−0.989010989	+0.989010989	+0.989010989
2	−0.010868253	+0.010868253	+0.010868253
3	−0.000119433	+0.000119433	+0.000119433
4	−0.000001311	+0.000001311	+0.000001311
5	−0.000000013	+0.000000013	+0.000000013
6	−0.000000002	+0.000000002	+0.000000002

It is seen that for all m's listed, the ratio

$$\left| \frac{\xi_M^{(m)}}{\xi_M^{(m-1)}} \right| < 1$$

Hence, it seems plausible that convergence will ensue. From the above table one has,

$$\begin{aligned} x_1 &\simeq \sum_{j=1}^{6} \xi_1^{(j)} = -1.000000001 \\ x_2 &\simeq \sum_{j=1}^{6} \xi_2^{(j)} = +1.000000001 \\ x_3 &\simeq \sum_{j=1}^{6} \xi_3^{(j)} = +1.000000001 \end{aligned}$$

The improvement in the accuracy of the solution over that obtained by the direct method of elimination is obvious. However, in this case, estimation of error becomes difficult.

An ill-conditioned set involving equations of ten unknowns has been successfully solved using the present method. This arises in the evaluation of sliding contact temperature distribution at the tool-chip interface. Readers are referred to References Cited[3] for details.

VI. REFERENCES CITED

1. A. M. Turing, "Rounding Off Errors in Matrix Process," *Quarterly Journal of Mechanics and Applied Mathematics,* Vol. 1 (1948), pp. 287-308.

2. J. W. Head and G. M. Oulton, "The Solution of 'Ill-Conditioned' Linear Simultaneous Equations," *Aircraft Engineering,* Vol. 30 (October, 1958), pp. 309-312.

3. B. T. Chao, H. L. Li, and K. J. Trigger, "Experimental Determination of Temperature Distribution at Tool Flank and Evaluation of Frictional Energy Distribution over Tool-Chip Contact," *ME Technical Report ORD-1980-5,* University of Illinois, March, 1958.

4. F. S. Shaw. *An Introduction to Relaxation Methods.* New York: Dover Publications, Inc., 1953. Pp. 20-25.

5. R. A. Buckingham. *Numerical Methods.* London: Sir Isaac Pitman and Sons, Ltd., 1957. Pp. 423-445, 533-534.

6. L. Fox, "A Short Account of Relaxation Methods," *Quarterly Journal of Mechanics and Applied Mathematics,* Vol. 1 (1948), pp. 253-280.

7. G. Temple, "Relaxation Methods Applied to Linear Systems," *Proceedings of Royal Society A. 169,* 1939, pp. 476-486.

8. O. Taussky, "Note on the Conditions of Matrices," *Math Tables and Aids to Computation,* Vol. 4, No. 30 (April, 1950), pp. 111-112.

9. K. S. Kunz. *Numerical Analysis.* New York: McGraw-Hill Book Company, Inc., 1957. Pp. 289-292, 314-317.

10. J. Morris, "A Successive Approximation Process for Solving Simultaneous Linear Equations," *Aeronautical Research Committee Report and Memorandum No. 1711,* 1936.

11. L. Fox, H. D. Huskey, and J. H. Wilkinson, "Notes on the Solution of Algebraic Linear Simultaneous Equations," *Quarterly Journal of Mechanics and Applied Mathematics,* Vol. 1 (1948), pp. 149-173.

12. E. M. Neville, "Ill-Conditioned Sets of Linear Equations," *Philosophical Magazine,* Vol. 39 (1948), pp. 35-48.

13. A. D. Booth. *Numerical Methods.* London: Butterworths Scientific Publications, 1957. Pp. 72-98.

14. H. Geiringer, "On the Solution of Systems of Linear Equations by Certain Iterative Methods," *Reissner Anniversary Volume,* 1949, pp. 365-393.

15. K. L. Nielsen. *Methods in Numerical Analysis.* New York: The MacMillan Company, 1956. Pp. 194-199.

16. E. Bodewig. *Matrix Calculus.* New York: Interscience Publishers, Inc. 1956. Pp. 117-119.

17. "Computer Library Routine No. 100," Digital Computer Laboratory, University of Illinois, 1959.

18. G. E. Forsythe, "Solving Linear Algebraic Equations Can Be Interesting," *Bulletin of the American Mathematical Society,* Vol. 59 (1953), pp. 299-329.

19. B. T. Chao, H. L. Li, and K. J. Trigger, "On the Solution of Ill-Conditioned, Algebraic, Linear, Simultaneous Equations by Machine Computation," *ME Technical Report ORD-1980-7,* University of Illinois, June, 1959.

20. J. von Neumann and H. H. Goldstine, "Numerical Inverting of Matrices of High Order," *American Mathematical Society Proceedings,* Vol. 53 (1947), pp. 1027-1099.

21. D. R. Hartree. *Numerical Analysis.* Oxford: Clarendon Press, 1955. Pp. 152-155.

22. J. D. Riley, "Solving Systems of Linear Equations with a Positive Definite, Symmetric, but Possibly Ill-Conditioned Matrix," *Math Tables and Other Aids to Computation,* Vol. 9, No. 51 (July, 1955), pp. 96-101.

The Engineering Experiment Station was established by act of the University of Illinois Board of Trustees on December 8, 1903. Its purpose is to conduct engineering investigations that are important to the industrial interests of the state.

The management of the Station is vested in an Executive Staff composed of the Dean of Engineering, the Director, the heads of the departments in the College of Engineering, the professor in charge of Chemical Engineering, and the Director of Engineering Information and Publications. This staff is responsible for establishing the general policies governing the work of the Station. All members of the College of Engineering teaching staff are encouraged to engage in the scientific research of the Station.

To make the results of its investigations available to the public, the Station publishes a series of bulletins. Occasionally it publishes circulars which may contain timely information compiled from various sources not readily accessible to the Station clientele or may contain important information obtained during the investigation of a particular research project but not having a direct bearing on it. A few reprints of articles appearing in the technical press and written by members of the staff are also published.

In ordering copies of these publications reference should be made to the Engineering Experiment Station Bulletin, Circular, or Reprint Series number which is at the upper left hand corner on the cover. Address

ENGINEERING PUBLICATIONS OFFICE
114 CIVIL ENGINEERING HALL
UNIVERSITY OF ILLINOIS
URBANA, ILLINOIS